ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΤΜΗΜΑ ΙΑΤΡΙΚΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ

‘‘ΜΕΘΟΔΟΛΟΓΙΑ ΒΙΟΙΑΤΡΙΚΗΣ ΕΡΕΥΝΑΣ, ΒΙΟΣΤΑΤΙΣΤΙΚΗ ΚΑΙ ΚΛΙΝΙΚΗ ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ’’

Comparison of different types of clustering

Σύγκριση μεθόδων της ανάλυσης κατά συστάδων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΧΑΛΙΤΣΙΟΣ ΧΡΗΣΤΟΣ

Τριμελής Επιτροπή: Στεφανίδης Ι.(επιβλέπων)

Δοξάνη Χρ.

Ζιντζαράς Ηλ.

Λάρισα, Σεπτέμβριος 2017

ΠΕΡΙΛΗΨΗ

Στην συγκεκριμένη διπλωματική εργασία αναλύεται το πρόβλημα της ανάλυσης συστάδων και συγκρίνονται οι τεχνικές και οι μέθοδοι αυτών. Σκοπός της ανάλυσης συστάδων είναι να ομαδοποιεί τα στοιχεία σε cluster έτσι ώστε τα στοιχεία που ανήκουν στο ίδιο cluster να έχουν μεγαλύτερη ομοιότητα από τα στοιχεία που ανήκουν σε διαφορετικά cluster.
Στο πρώτο κεφάλαιο της εργασίας παρουσιάζεται η βασική ιδέα της ανάλυσης συστάδων, η γεωμετρική της ερμηνεία βασισμένη σε ένα παράδειγμα και τέλος ο σκοπός της ανάλυσης συστάδων. Εν συνεχεία παρουσιάζονται οι μέθοδοι που διαιρούνται σε ιεραρχικές μεθόδου και μη ιεραρχικές μεθόδους. Λόγω των πολλών διαφορετικών τρόπων προσδιορισμού των cluster υπάρχουν πολλές διαφορετικές τεχνικές ανάλυσης συστάδων που αντιστοιχούν σε κάθε μέθοδο και διαφέρουν κυρίως σε σχέση με το πώς υπολογίζονται οι αποστάσεις μεταξύ των cluster.
Παρουσιάζονται αναλυτικά οι τεχνικές ιεραρχικής ομαδοποίησης όπως απλής σύνδεσης, πλήρης σύνδεσης κα σύνδεσης μέσου όρου τονίζοντας βασικές ιδιότητες και διαφορές μεταξύ αυτών. Επίσης αναφέρονται βασικά πλεονεκτήματα και μειονεκτήματα της ιεραρχικής μεθόδου. Στην συνέχεια αναλύονται οι μη ιεραρχικές μέθοδοι με σημαντικότερο αλγόριθμο τον K-means με τα πλεονεκτήματα και μειονεκτήματά του. Οι μη ιεραρχικές μέθοδοι ομαδοποίησης χρησιμοποιούνται κυρίως για να βελτιώνουν την λύση των cluster που έχει προκύψει από μια ιεραρχική μέθοδο. Στο σημείο αυτό της εργασίας γίνεται μια σύνοψη των διαφορών των αλγορίθμων ομαδοποίησης και σύγκρισης τους από εμπειρικές μελέτες.

ABSTRACT

In the current diplomatic project is analyzed the problem of cluster analysis and be compared the different types of it. The purpose of cluster analysis is to group items in cluster, so that items belonging to the same cluster have a greater similarity than the items belonging to different clusters.

In the first chapter of the project, the basic idea of cluster analysis is presented, as well as the geometric interpretation of it which is based on a simple instance. Then, cluster methods are presented which are divided into hierarchical and non-hierarchical methods or optimization method. Because of the different ways determining the clusters, there are many different techniques of cluster analysis, associated with each method and they differ mainly, compared with the calculation of distances between clusters.

The hierarchical clustering techniques are presented analytically such as single linkage, complete linkage and average linkage emphasizing their differences and properties. Additionally, advantages and drawbacks of hierarchical methods are mentioned. Afterwards, the non-hierarchical methods are analyzed with more important K-means. Non-hierarchical methods are used mainly to improve the cluster solution which has resulted from a hierarchical method. In this part of thesis, a summary of the various clustering algorithms is presented with their comparison from empirical studies.

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΕΦΑΛΑΙΟ 1

ΕΙΣΑΓΩΓΗ

1.1 Ανάλυση Συστάδων – Ορισμός

Η ανάλυση συστάδων ή clustering, είναι η οργάνωση μιας συλλογής από σημεία σε ένα n-διάστατο ή διανύσματα χώρο, σε συστάδες (clusters), με βάση κάποιο μέτρο ομοιότητας. Στοιχεία που ανήκουν στην ίδια ομάδα, παρουσιάζουν μεγαλύτερη ομοιότητα, από τα στοιχεία που ανήκουν σε διαφορετικές ομάδες.
Σε διάφορες επιστημονικές έρευνες, ο ερευνητής ενδιαφέρεται να βρει μια ταξινόμηση στην οποία τα αντικείμενα που τον ενδιαφέρουν, ταξινομούνται σε ένα μικρό αριθμό ομοιογενών ομάδων ή clusters. Συχνά αυτήν η ταξινόμηση εξυπηρετεί θεμελιώδεις σκοπούς. Στην ψυχιατρική, η ταξινόμηση των ψυχικών διαταραχών, θα βοηθούσε στην διερεύνηση των αιτιών τους και θα οδηγούσε σε πιο βελτιωμένες μεθόδους θεραπείας. Επιπλέον, η ομαδοποίηση γονιδίων που έχουν την ίδια λειτουργία είναι κορυφαίας σημασίας.
Η ομαδοποίηση ή clustering, διακρίνεται από τις μεθόδους ταξινόμησης. Η ανάλυση συστάδων είναι μια τεχνική στην οποία δεν γίνεται καμία υπόθεση σχετικά με τον αριθμό των ομάδων ή τη δομή της ομάδας. Το γεγονός ότι δεν υπάρχει μια εκ των προτέρων ταξινόμηση του δείγματος υποδηλώνει ότι η ανάλυση συστάδων είναι θεμελιώδης εργαλείο για την εξερεύνηση των δεδομένων. Έτσι, το πρώτο πράγμα που πρέπει να σημειωθεί είναι ότι δεν οδηγούμαστε πάντα στην ίδια ταξινόμηση και πάντα θα υπάρχει μια ποικιλία από εναλλακτικές ταξινομήσεις για το ίδιο σύνολο αντικειμένων ή ατόμων. Για παράδειγμα, τα ανθρώπινα όντα μπορούν να ταξινομηθούν βάσει το φύλο τους, την ηλικία, ή ακόμη και το μορφωτικό τους επίπεδο. Γενικά, το είδος της ομαδοποίησης που προκύπτει από μια ανάλυση εξαρτάται από τις μεταβλητές που αναπαριστούν το αντικείμενο.

1.2 Ευκλείδεια Απόστασης για Ζεύγη Στοιχείων και Πίνακας Απόστασης

Το πιο γνωστό από τα μέτρα ομοιότητας είναι η Ευκλείδεια απόσταση για 2 αντικείμενα x και y και ισούται με :

$$d(x, y)=\sqrt{\sum_{i=1}^{k}(xi - yi)^2}.$$

Η ευκλείδεια απόσταση χρησιμοποιείται ευρέως σε περιπτώσεις λίγων διαστάσεων και έχει καλά αποτελέσματα αν τα δεδομένα ομαδοποιούνται σε συμπαγή και αρκετά απομονωμένα cluster. Ωστόσο ένα πρόβλημα κατά την χρήσης σε πολλές διαστάσεις είναι πως το χαρακτηριστικό με την μεγαλύτερη διαφοροποίηση σε σχέση με τα υπόλοιπα κυριαρχεί και αποπροσανατολίζει το αποτέλεσμα.

Για 2 n-διάστατες παρατηρήσεις : x=[x1,x2,….,xn] και y=[y1,y2,….,yn] η Ευκλείδεια απόσταση δίνεται από τον τύπο:

$$d(x, y)=\sqrt{(x1 - y1)^2 + (x2 - y2)^2 + \cdots + (xn - yn)^2}.$$

Για παράδειγμά έστω ότι θέλουμε να υπολογίσουμε την ευκλείδεια απόσταση μεταξύ 2 οποιοδήποτε ασθενών από των παρακάτω πίνακα και να δημιουργήσω τον πίνακα αποστάσεων.

patient	**X1**	**X2**	**X3**
1	x11=22	x21=21	x13=28
2	x21=20	x22=22	x23=30
3	x31=14	x32=15	x33=21
4	x41=16	x42=16	x43=24
5	x51=19	x52=19	x53=26

Τότε η απόσταση μεταξύ ασθενή 1 και 2 είναι:

$d(1,2)=\sqrt{(x11 - x21)^2 + (x12 - x22)^2 + (x13 - x23)^2}$ =

$\sqrt{(22 - 20)^2 + (21 - 22)^2 + (28 - 30)^2}$ = 3. Όμοια υπολογίζονται και οι υπόλοιπες διαφορές και παίρνω τον πίνακα αποστάσεων:

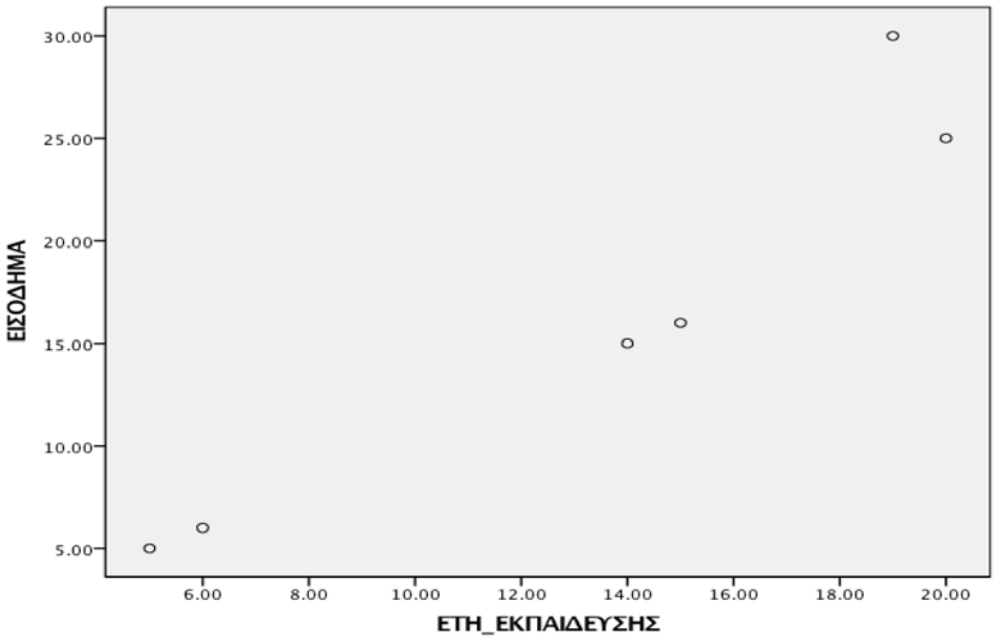
patient	**1**	**2**	**3**	**4**	**5**
1	0				
2	3	0			
3	12.2	12.9	0		
4	8.77	9.38	3.74	0	
5	4.9	5.4	7.55	4.12	0

1.3 Γεωμετρική Ερμηνεία της Ανάλυσης Συστάδων

Η γεωμετρική της ερμηνεία είναι πολύ απλή. Θεωρώ τα υποθετικά δεδομένα του παρακάτω πίνακα:

ΑΝΤΙΚΕΙΜΕΝΟ	*ΕΙΣΟΔΗΜΑ*	*ΕΤΗ ΕΚΠΑΙΔΕΥΣΗΣ*
S1	5	5
S2	6	6
S3	15	14
S4	16	15
S5	25	20
S6	30	19

Όπως φαίνεται στο παρακάτω σχήμα κάθε παρατήρηση αναπαρίσταται σε ένα σημείο του δισδιάστατου χώρου. Γενικά, κάθε παρατήρηση μπορεί να αναπαρασταθεί σαν ένα σημείο σε ένα n- διάστατο χώρο, όπου n ο αριθμός των χαρακτηριστικών ή μεταβλητών που χρησιμοποιούνται για να περιγράψουμε τα αντικείμενα.
Υποθέτω ότι θέλω να δημιουργήσω 3 ομοιογενείς ομάδες. Σύμφωνα με το σχήμα τα S1, S2 τα S3, S4 και τα S5, S6 αποτελούν 3 ομάδες.



Όπως βλέπουμε η ανάλυση συστάδων ομαδοποιεί παρατηρήσεις ώστε οι παρατηρήσεις σε κάθε ομάδα να είναι όμοιες σε σχέση με τις μεταβλητές ομαδοποίησης.

1.4 Σκοπός Ανάλυσης Συστάδων

Σκοπός της ανάλυσης συστάδων είναι να ομαδοποιεί σε cluster έτσι ώστε κάθε cluster να είναι όσο το δυνατό ομοιογενές σε σχέση με τις μεταβλητές ομαδοποίησης. Το πρώτο βήμα είναι να επιλέξουμε ένα μέτρο ομοιότητας με το οποίο μετράμε την συσχέτιση(ομοιότητα) μεταξύ των αντικειμένων. Στην συνέχεια εξετάζουμε το είδος της τεχνικής ομαδοποίησης που θα χρησιμοποιήσουμε(ιεραρχική ή μη ιεραρχική). Τρίτο βήμα είναι να επιλεγεί το είδος της μεθόδου για την επιλεγμένη τεχνική. Και τέλος, γίνεται μια συζήτηση αναφορικά με τον αριθμό των cluster και ερμηνεύονται τα αποτελέσματα.

Στην πορεία της διπλωματικής θα ασχοληθούμε με τις δύο παρακάτω κατηγορίες τεχνικών που αποτελούν την πλειονότητα των εφαρμογών στην ανάλυση συστάδων.

1) Συσσωρευτικοί Ιεραρχικοί μέθοδοι (single linkage, complete linkage, average linkage)
2) Μη ιεραρχικοί μέθοδοι (k-means)

ΚΕΦΑΛΑΙΟ 2

ΙΕΡΑΡΧΙΚΕΣ ΜΕΘΟΔΟΙ

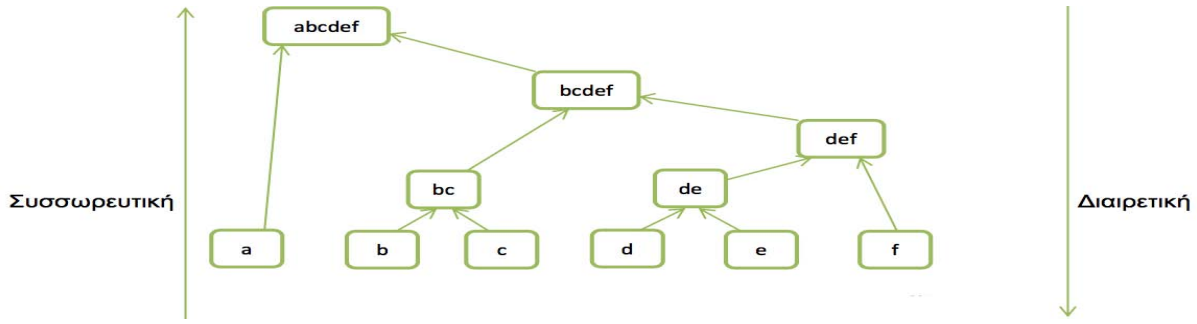
Οι ιεραρχικές τεχνικές ομαδοποίησης είτε με μια σειρά διαδοχικών συγχωνεύσεων είτε με μια σειρά διαδοχικών διαιρέσεων και γι' αυτό το λόγο χωρίζονται σε συσσωρευτικές και διαιρετικές.

Οι **συσσωρευτικές(agglomerative)** ιεραρχικές μέθοδοι ξεκινούν μεμονωμένα αντικείμενα. Κατά συνέπεια υπάρχουν αρχικά τόσες πολλές συστάδες όσες είναι και τα αντικείμενα όντας κάθε αντικείμενο μια ξεχωριστή συστάδα. Τα πιο όμοια αντικείμενα ομαδοποιούνται πρώτα και οι αρχικές αυτές ομάδες συγχωνεύονται σύμφωνα με τις ομοιότητές τους στο επόμενο υψηλότερο επίπεδο με ένα cluster λιγότερο. Το ζευγάρι που επιλέχθηκε για την συγχώνευση αποτελείται από 2 ομάδες με την μικρότερη ανομοιότητα μέσα στην ομάδα. Τελικά καθώς μειώνεται η ομοιότητα όλες οι υποομάδες συγχωνεύονται σε μια ενιαία συστάδα.

Οι **διαιρετικές(divisive)** ιεραρχικές μέθοδοι δουλεύουν στην αντίθετη κατεύθυνση. Μια αρχική ενιαία ομάδα διαιρείται σε 2 υποομάδες. Η διάσπαση επιλέγεται για να παράγει 2 νέες ομάδες με την μεγαλύτερη ανομοιότητα μεταξύ των ομάδων. Αυτές οι υποομάδες διαιρούνται περαιτέρω σε ανόμοιες υποομάδες. Η διαδικασία συνεχίζεται έως ότου υπάρξουν τόσες πολλές υποομάδες όσες και τα αντικείμενα, δηλαδή μέχρι κάθε αντικείμενο να σχηματίσει μια ομάδα.

Γι' αυτό και agglomerative είναι η προσέγγιση από κάτω προς τα πάνω αφού αρχικά κάθε αντικείμενο είναι ένα cluster, κι όσο ανεβαίνουμε προς τα πάνω στην ιεραρχία ζευγαρώνει με άλλα cluster. Ενώ divisive είναι η προσέγγιση από πάνω προς τα κάτω.

Στο παρακάτω παράδειγμα παρουσιάζονται οι 2 προσεγγίσεις στο οποίο θέλουμε να ομαδοποιήσουμε 6 cluster: a, b, c, d, e, f.



2.1 Συσσωρευτικές Ιεραρχικές Μέθοδοι

Σε μια ιεραρχική ταξινόμηση τα δεδομένα δεν διαμερίζονται σε ένα συγκεκριμένο αριθμό κατηγοριών ή συστάδων σε ένα βήμα μόνο. Αντί γι' αυτό, η ταξινόμηση αποτελείται από μια σειρά διαμερίσεων, που εκτελούνται από μια μοναδική συστάδα , που περιέχει όλα τα άτομα σε n συστάδες που η κάθε μια περιέχει ένα άτομο. Οι συσσωρευτικές ιεραρχικές τεχνικές ομαδοποίησης παράγουν διαμερίσεις από μια σειρά διαδοχικών συγχωνεύσεων των n ατόμων σε ομάδες. Οι συγχωνεύσεις είναι μη αναστρέψιμες, δηλαδή όταν ένας συσσωρευτικός αλγόριθμος έχει θέσει 2 άτομα στην ίδια ομάδα δεν μπορούν αυτά να εμφανιστούν στην συνέχεια σε διαφορετικές ομάδες.

Υποθέτουμε ότι μια συσσωρευτική μέθοδος έχει φτάσει στο στάδιο που έχει c cluster. Το επόμενο βήμα είναι να συγχωνεύσουμε 2 από αυτά σε 1 για να παράγουμε c-1 clusters. Αυτό επαναλαμβάνεται στην συνέχεια, για να δώσει c-2 clusters, και ούτω καθεξής. Φυσικά, αρχικά c=n ο αριθμός των σημείων παρατήρησης. Σε κάθε βήμα τα 2 clusters που συγχωνεύονται επιλέγονται μελετώντας τον πίνακα απόστασης των αποστάσεων μεταξύ των cluster. Οι υποψήφιοι για την συγχώνευση σε κάθε στάδιο είναι τα 2 πλησιέστερα cluster, με τους τρόπους μέτρησης των αποστάσεων να διαφέρουν.

Όταν όλες συσσωρευτικές τεχνικές περιορίσουν τελικά τα δεδομένα σε μια μόνο συστάδα που περιέχει όλα τα άτομα ο ερευνητής που επιδιώκει την λύση με την καλύτερη ΄΄προσαρμογή΄΄του αριθμού των συστάδων θα αποφασίσει ποια διαμέριση να επιλέξει.

Ένας λόγος ύπαρξης διαφορετικών τεχνικών ανάλυσης είναι ότι υπάρχουν πολλοί διαφορετικοί τρόποι προσδιορισμού των cluster. Οι διάφοροι μέθοδοι διαφέρουν ουσιαστικά στον τρόπο υπολογισμού της απόστασης ανάμεσα στα 2 cluster. Οι πιο δημοφιλείς μέθοδοι είναι:

1) Κοντινότερος γείτονας ή single linkage
2) Μακρινότερος γείτονας ή complete linkage
3) Σύνδεση μέσου όρου ή average linkage
4) Σύνδεση κεντροειδούς ή centroid
5) Μέθοδος Ward
6) Σύνδεση διαμέσου

Εμείς θα ασχοληθούμε στην παρούσα διπλωματική με την παρουσίαση και σύγκριση των 3 πρώτων μεθόδων.

2.2 Συσσωρευτικός Ιεραρχικός αλγόριθμος

Ο συσσωρευτικός ιεραρχικός αλγόριθμος περιλαμβάνει τα εξής βήματα:

Βήμα 1: Ξεκινώ με N συστάδες που κάθε μια περιέχει μόνο μια οντότητα και έναν NxN πίνακα αποστάσεων συμμετρικό D={d(ik)}.

Βήμα 2: Αναζητώ στον πίνακα αποστάσεων το κοντινότερο(πιο όμοιο) ζευγάρι συστάδων. Έστω ότι η απόσταση στα cluster U-V είναι d(UV).

Βήμα 3: Συγχωνεύομε τα cluster U-V και ονομάζω την πιο πρόσφατη σχηματισμένη συστάδα (UV). Επαναυπολογίζουμε τις εισόδους στον πίνακα αποστάσεων διαγράφοντας τις στήλες και τις γραμμές που αντιστοιχούν στα cluster U και V και προσθέτοντας μια γραμμή και στήλη που δίνει τις αποστάσεις ανάμεσα στην συστάδα (UV) και τις υπόλοιπες.

Βήμα 4: Επαναλαμβάνουμε τα βήματα 2 και 3 συνολικά N-1 φορές και καταγράφουμε τις συστάδες που συγχωνεύθηκαν και τα επίπεδα (αποστάσεις ή ομοιότητες) στα οποία πραγματοποιούνται οι συγχωνεύσεις.

2.3 Μέθοδος Απλής Σύνδεσης (single linkage)

Είπαμε ότι οι συσσωρευτικές ιεραρχικές τεχνικές διαφέρουν κυρίως ως προς το πως μετρούν την απόσταση ανάμεσα σε 2 cluster-το οποίο cluster κάποιες φορές μπορεί να αποτελείται από μόνο ένα άτομο-. Δύο απλά μέτρα μεταξύ των ομάδων είναι:

d(AB) = min {d(i, j)} (1)

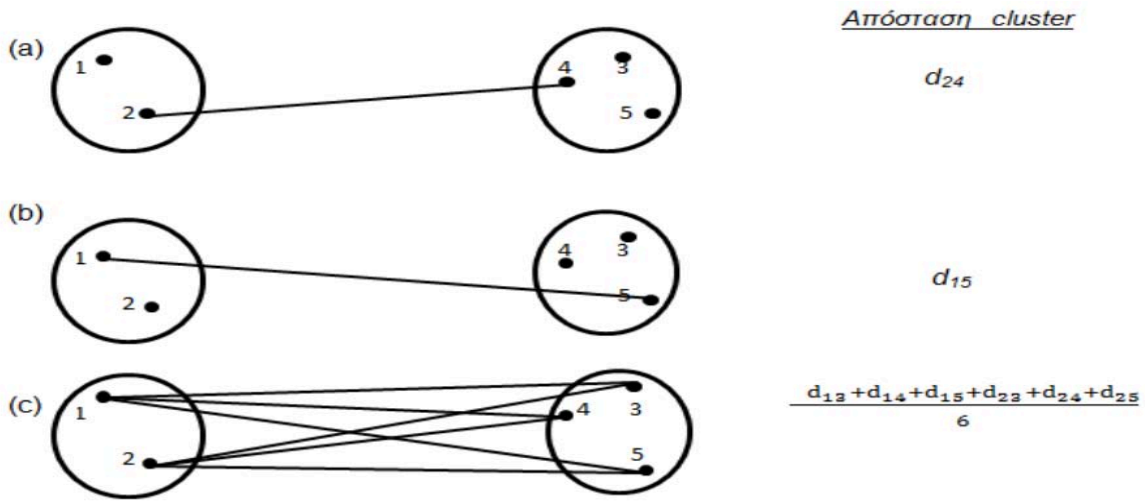
d(AB) = max {d(i, j)} (2), i, j ∈ A, B

όπου d(AB) η απόσταση ανάμεσα σε 2 cluster A και B και d(i, j) η απόσταση ανάμεσα στα άτομα i, j. Αυτήν θα μπορούσε να είναι η ευκλείδεια απόσταση ή κάποιο άλλο μέτρο απόστασης.

Το μέτρο ανομοιότητας μεταξύ ομάδων (1) είναι η βάση της ομαδοποίησης απλής σύνδεσης ενώ το (2) της ομαδοποίησης πλήρους σύνδεσης. Μια περαιτέρω δυνατότητα για την μέτρηση της απόστασης μεταξύ των cluster είναι:

$$d(AB) = \frac{1}{n(A)n(B)} \sum i \in A \sum j \in B \; d(i, j)$$

όπου n(A), n(B) οι αριθμοί ατόμων στα 2 cluster A και B. Το μέτρο αυτό είναι η βάση της average linkage. Σχηματικά τα 3 παραπάνω μέτρα στο ακόλουθο σχήμα για (a)απλή σύνδεση,(b) πλήρη σύνδεση,(c) σύνδεση μέσου όρου:



Η **single linkage** αποκαλείται και ως η μέθοδος του κοντινότερου γείτονα. Οι είσοδοι στον αλγόριθμος της μπορεί να είναι αποστάσεις ή ομοιότητες μεταξύ ζευγών αντικειμένων. Οι ομάδες σχηματίζονται από μεμονωμένες οντότητες συγχωνεύοντας τους κοντινότερους γείτονες(την μικρότερη απόσταση ή την μεγαλύτερη ομοιότητα).
Εφαρμόζοντάς την έχουμε τον πίνακα D = {d(i, k)} και πρέπει να βρούμε την μικρότερη απόσταση που υπάρχει σε αυτόν και στην συνέχεια συγχωνεύουμε τα αντίστοιχα αντικείμενα με την μικρότερη απόσταση τα λεγόμενα U,V για να πάρουμε την συστάδα (UV). Για το βήμα 3 του γενικού αλγορίθμου οι αποστάσεις μεταξύ (UV) και κάποιας W είναι: d{(UV), W} = min {d(UW, VW)}, εννοείται η κοντινότερες αποστάσεις.

2.3.1 Ομαδοποίηση με τη Χρήση Απλής Σύνδεσης

Για να γίνει καλύτερα κατανοητή η λειτουργεία της απλής σύνδεσης θεωρώ υποθετικές αποστάσεις μεταξύ 5 αντικειμένων:

D = {d(i, k)} =	(1)	(2)	(3)	(4)	(5)
(1)	0				
(2)	9	0			
(3)	3	7	0		
(4)	6	5	9	0	
(5)	11	10	2	8	0

Θεωρώντας κάθε αντικείμενο ως συστάδα η ομαδοποίηση ξεκινά με την συγχώνευση των 2 κοντινότερων στοιχείων, δηλαδή των 5 και 3 αφού min {d(i, k)}= 2 σχηματίζοντας την συστάδα (35).

Για το επόμενο στάδιο χρειαζόμαστε τις αποστάσεις μεταξύ (35) και των 1,2,4. Έτσι, d{(35), 1}= min {d(31, 51)}= 3, d{(35), 2}= min {d(32, 52)}= 7, d{(35), 4}= min {d(34, 54)}= 8. Διαγράφω γραμμές και στήλες του D που αντιστοιχούν στα αντικείμενα 3 και 5 και προσθέτοντας μια γραμμή και στήλη για την συστάδα (35) έχω:

	(35)	(1)	(2)	(4)
(35)	0			
(1)	3	0		
(2)	7	9	0	
(4)	8	6	5	0

Η μικρότερη απόσταση είναι μεταξύ (35), 1 με d{(35), 1}=3 οπότε συγχωνεύουμε τις συστάδες και προκύπτει η (135). Υπολογίζοντας τις νέες αποστάσεις έχω, d{(135), 2}= min {d((35)2,12}= 7, d{(135), 4}= min {d((35)4, 41)}= 6 και ο πίνακας είναι:

	(35)	(2)	(4)
(135)	0		
(2)	7	0	
(4)	6	5	0

Η μικρότερη απόσταση είναι μεταξύ 4,2 με d{4,2}=5 οπότε συγχωνεύουμε τις συστάδες και προκύπτει η (24). Έτσι καταλήξαμε να έχουμε δύο ευδιάκριτες συστάδες τις (135), (24). Υπολογίζοντας την απόσταση του κοντινότερου γείτονά τους προκύπτει, d{(135), (24)}= min {d((135)2,(135)4}= 6 και ο τελικός πίνακας είναι

	(35)	(2)
(135)	0	
(24)	6	0

Συνεπώς τα cluster συγχωνεύονται για να σχηματίσουν μια ενιαία συστάδα όλων των αντικειμένων την (12345) όταν η απόσταση του κοντινότερου γείτονα φτάσει στο 6.

Οι διαμερίσεις κάθε σταδίου είναι οι εξής:

Στάδιο	Ομάδες
P5	[1],[2],[3],[4],[5]
P4	[35],[1],[2],[4]
P3	[135],[2],[4]
P2	[135],[24]
P1	[12345]

Ένα σύνηθες πρόβλημα της απλής σύνδεσης είναι συνενώνει συστάδες οι οποίες έχουν δύο κοντινά σημεία και πολλά άλλα σημεία βρίσκονται σε μεγάλες αποστάσεις. Ένα άλλο πρόβλημα είναι ότι μπορεί να προκληθεί η δημιουργία μιας επιμήκους συστάδας και να προστίθενται συνεχώς νέα σημεία στην ουρά της συστάδας. Επίσης εάν μεταξύ δύο πραγματικών συστάδων υπάρχουν μεμονωμένα σημεία που δημιουργούν μια γέφυρα τότε οι συστάδες θα ενωθούν, έτσι τα σημεία στα άκρα της συστάδας θα απέχουν πολύ μεταξύ τους. Το πρόβλημα αυτό είναι γνωστό ως φαινόμενο της αλυσίδας(chaining phenomenon). Το σημαντικό πλεονέκτημά της είναι ότι δεν επηρεάζεται από ακραίες τιμές.

2.4 Μέθοδος Πλήρης Σύνδεσης (complete linkage)

Η ομαδοποίηση πλήρους σύνδεσης προχωρά με τον ίδιο τρόπο με της απλής σύνδεσης αλλά με μια σημαντική εξαίρεση. Σε κάθε στάδιο η απόσταση ανάμεσα στις συστάδες καθορίζεται από την απόσταση ανάμεσα σε 2 στοιχεία, ένα από κάθε συστάδα, που είναι πιο απόμακρα και γι’ αυτό καλείται και η μέθοδος του μακρινότερου γείτονα.
Ο γενικός συσσωρευτικός αλγόριθμος, αρχίζει ξανά, βρίσκοντας την ελάχιστη είσοδο στον πίνακα D = {d(i, k)} και συγχωνεύοντας αντίστοιχα τα αντικείμενα U,V για να πάρουμε την συστάδα (UV). Στο βήμα 3 του γενικού αλγορίθμου οι αποστάσεις μεταξύ (UV) και W υπολογίζεται από d{(UV), W} = max {d(UW, VW)}.

2.4.1 Ομαδοποίηση με την Χρήση Πλήρους Σύνδεσης

Θεωρώ ξανά τον πίνακα αποστάσεων:

D = {d(i, k)}=

	(1)	(2)	(3)	(4)	(5)
(1)	0				
(2)	9	0			
(3)	3	7	0		
(4)	6	5	9	0	
(5)	11	10	2	8	0

Στο πρώτο στάδιο τα αντικείμενα 3 και 5 συγχωνεύονται αφού είναι τα δύο πιο όμοια, σχηματίζοντας την συστάδα (35). Στο στάδιο δύο υπολογίζουμε:
d{(35), 1}= max {d(31, 51)}= 11, d{(35), 2}= max {d(32, 52)}= 10, d{(35), 4}= max {d(34, 54)}= 9 και ο νέος πίνακας είναι:

	(35)	(1)	(2)	(4)
(35)	0			
(1)	11	0		
(2)	10	9	0	
(4)	9	6	5	0

Η επόμενη συγχώνευση συμβαίνει μεταξύ των δύο πιο όμοιων ομάδων 2 και 4 γι να πάρουμε την συστάδα (24). Στο βήμα 3 έχουμε:

d{(35), (35)}= max {d((2,(35)), (4,(35)))}= 10, d{(24), 1}= max {d(2, 1), (4, 1) }= 9
και ο πίνακας αποστάσεων είναι:

	(35)	(24)	(1)
(35)	0		
(24)	10	0	
(1)	11	9	0

Η επόμενη συγχώνευση παράγει την συστάδα (124) και στο τελικό στάδιο συγχωνεύονται οι (35) και (124) σε μια ενιαία συστάδα (12345) στο επίπεδο με:

	(124)	(35)
(124)	0	
(35)	11	0

Συνεπώς τα cluster συγχωνεύονται για να σχηματίσουν μια ενιαία συστάδα όλων των αντικειμένων την (12345) όταν η απόσταση φτάσει στο 6.
Οι διαμερίσεις κάθε σταδίου είναι οι εξής:

Στάδιο	Ομάδες
P5	[1],[2],[3],[4],[5]
P4	[35],[1],[2],[4]
P3	[35],[1],[24]
P2	[35],[124]
P1	[12345]

Η μέθοδος δεν συνίσταται για δεδομένα στα οποία μπορεί να υφίσταται αρκετός θόρυβος (θόρυβος είναι τυχαίο σφάλμα ή διακύμανση σε μια μετρούμενη μεταβλητή , Jiawei Han, Micheline Kamber, Jian Pei, 2012).
Με την μέθοδο αυτήν αποφεύγονται προβλήματα που παρουσιάζονται με την απλή σύνδεση όπως επιμήκεις συστάδες. Αντίθετα η πλήρης σύνδεση τείνει να δημιουργεί συμπαγείς συστάδες και θεωρείται χρήσιμη αν είναι αναμενόμενο ότι οντότητες της ίδιας συστάδας βρίσκονται σε μεγάλη μεταξύ τους απόσταση στον πολυδιάστατο χώρο.

2.5 Μέθοδος Σύνδεση Μέσου Όρου (average linkage)

Η σύνδεση του μέσου όρου θεωρεί την απόσταση μεταξύ 2 cluster σαν την μέση απόσταση ανάμεσα σ όλα τα ζευγάρια στοιχείων, όπου ένα μέλος κάθε ζευγαριού ανήκει σε κάθε cluster (δηλαδή το μέσο της απόστασης των μελών του ενός cluster και του άλλου).
Ξανά οι είσοδοι στον αλγόριθμο μπορεί να είναι αποστάσεις και η μέθοδος χρησιμεύει για να ομαδοποιεί αντικείμενα ή μεταβλητές. Σύμφωνα με τον γενικό αλγόριθμο ξεκινώ ψάχνοντας τον πίνακα απόστασης D για να βρούμε τα πλησιέστερα(πιο όμοια) αντικείμενα, για παράδειγμα τα U,V και τα συγχωνεύουμε για να προκύψει η συστάδα (UV). Στο 3^{o} βήμα οι αποστάσεις μεταξύ (UV) και W καθορίζονται από τον τύπο d((UV)W) = $\frac{1}{n(UV)n(W)} \sum i \in (UV) \sum k \in W$ d(i, k) με d(i, k) η απόσταση ανάμεσα στο αντικείμενο i και k.

2.5.1 Ομαδοποίηση με τη Σύνδεση του Μέσου Όρου

Θεωρώ ξανά τον πίνακα τον πίνακα αποστάσεων:

D = {d(i, k)}=	(1)	(2)	(3)	(4)	(5)
(1)	0				
(2)	9	0			
(3)	3	7	0		
(4)	6	5	9	0	
(5)	11	10	2	8	0

Στο πρώτο στάδιο τα αντικείμενα 3 και 5 συγχωνεύονται αφού είναι τα δύο πιο όμοια, σχηματίζοντας την συστάδα (35). Στο στάδιο δύο υπολογίζουμε:

d{(35), 1}=$\frac{d(3,1)+d(5,1)}{2}$=(3+11)/2=7, d{(35), 2}=$\frac{d(3,2)+d(5,2)}{2}$=(7+10)/2=8.5, d{(35), 4}=$\frac{d(3,4)+d(5,4)}{2}$=(9+8)/2=8.5 και ο νέος πίνακας είναι:

	(35)	(1)	(2)	(4)
(35)	0			
(1)	17	0		
(2)	8.5	9	0	
(4)	8.5	6	5	0

Η επόμενη συγχώνευση συμβαίνει μεταξύ των δύο πιο όμοιων ομάδων 2 και 4 γι να πάρουμε την συστάδα (24). Στο βήμα 3 έχουμε:

d{(24), (35)}=$\frac{d(2,35)+d(4,35)}{2}$=(8.5+8.5)/2=8.5, d{(24), (1)}=$\frac{d(2,1)+d(4,1)}{2}$=(9+6)/2=7.5 και ο νέος πίνακας είναι:

	(35)	(24)	(1)
(35)	0		
(24)	8.5	0	
(1)	11	7.5	0

Η επόμενη συγχώνευση παράγει την συστάδα (124) και d{(124), (35)}=$\frac{d(1,35)+d(24,35)}{2}$=(11+8.5)/2=9.75 άρα:

	(124)	(35)
(124)	0	
(35)	9.75	0

Συνεπώς τα cluster συγχωνεύονται για να σχηματίσουν μια ενιαία συστάδα όλων των αντικειμένων την (12345) όταν η απόσταση φτάσει στο 9.75.

Οι διαμερίσεις κάθε σταδίου είναι οι εξής:

Στάδιο	Ομάδες
P5	[1],[2],[3],[4],[5]
P4	[35],[1],[2],[4]
P3	[35],[1],[24]
P2	[35],[124]
P1	[12345]

Αυτή η μέθοδος απαιτεί το μεγαλύτερο κόστος σε υπολογισμούς καθώς υπολογίζει την μέση απόσταση όλων των πιθανών ζευγών στοιχείων από τις 2 συστάδες που διερευνώνται. Η χρήση της μεθόδου δεν δημιούργει το φαινόμενο της αλυσίδας (single linkage) ενώ τα απομακρυσμένα outliers δεν χρήζουν ιδιαίτερης σημασίας κατά την απόφαση δημιουργίας συστάδων. Άμεση συνέπεια αυτού είναι ότι η συγκεκριμένη μέθοδος είναι πιο δημοφιλής από τις άλλες. Επίσης η κάθε συσσώρευση συμβαίνει σε απόσταση συστάδων μεγαλύτερη από αυτήν του προηγούμενου σταδίου συσσώρευσης δίνοντας την δυνατότητα τερματισμού της ομαδοποίησης είτε όταν οι συστάδες βρίσκονται πολύ μακριά για να συγχωνευθούν, είτε όταν υπάρχει ικανοποιητικά μικρός αριθμός υπολογισμένων συστάδων.

2.6 Πλεονεκτήματα και Μειονεκτήματα Ιεραρχικών Μεθόδων

Τα βασικά πλεονεκτήματα είναι τα εξής:

- Δεν απαιτείται ο αριθμός των ομάδων να είναι γνωστός εκ των προτέρων.
- Δημιουργούν πολλαπλά επίπεδα φωλιασμένων συστάδων και επιτρέπουν στον χρήστη να επιλέξει το επίπεδο που αυτός επιθυμεί.
- Δεν υπάρχουν παράμετροι εισόδου (εκτός από την επιλογή της ομοιότητας)

Στον αντίποδα τα βασικά μειονεκτήματά του είναι τα ακόλουθα:

- Κάθε ενέργεια η οποία πραγματοποιείται σε ένα στάδιο δεν είναι αντιστρέψιμη. Από τη στιγμή που 2 αντικείμενα ενταχθούν στην ίδια ομάδα θα παραμείνουν στην ίδια ομάδα χωρίς την δυνατότητα να διαχωριστούν στην πορεία.
- Χρειάζεται να ελέγξουν πολλές αποστάσεις και έτσι καθυστερούν όταν χρειάζεται να επεξεργαστούν μεγάλο αριθμό αντικειμένων.

ΚΕΦΑΛΑΙΟ 3

ΜΗ ΙΕΡΑΡΧΙΚΕΣ ΜΕΘΟΔΟΙ

Στην μη ιεραρχική ομαδοποίηση τα δεδομένα διαιρούνται σε k διαμερίσεις ή ομάδες με κάθε διαμέριση να αντιπροσωπεύει ένα cluster. Σε αντίθεση με την ιεραρχική, ο αριθμός των cluster k, μπορεί είτε να διευκρινιστεί εκ των προτέρων είτε να καθοριστεί σαν μέρος της διαδικασίας ομαδοποίησης. Επειδή δεν είναι απαραίτητο να καθοριστεί ένας πίνακας αποστάσεων μεταξύ των αντικειμένων που θέλουμε να οργανώσουμε σε συστάδες, δεν είναι απαραίτητο να αποθηκευτούν στον υπολογιστή κατά το τρέξιμο του αλγορίθμου. Έτσι οι μη ιεραρχικές μπορούν να εφαρμοστούν σε πολύ μεγαλύτερο όγκο δεδομένων από τις ιεραρχικές.

Οι τεχνικές μη ιεραρχικής ομαδοποίησης ακολουθούν τα ακόλουθα βήματα:

Βήμα 1: Επιλέγουμε k αρχικά κεντροειδή των cluster ή κομβικά σημεία (seed points) με k να είναι ο επιθυμητός αριθμός cluster.

Βήμα 2: Αναθέτουμε κάθε παρατήρηση στο cluster στο οποίο αυτήν είναι η πλησιέστερη.

Βήμα 3: Αναθέτουμε εκ νέου ή ανακατεύουμε κάθε παρατήρηση σ ένα από τα k cluster σύμφωνα με ένα προκαθορισμένο κανόνα τερματισμού.

Βήμα 4: Σταματάει αν δεν υπάρχει καμία ανακατανομή των σημείων ή αν η ανακατανομή ικανοποιεί τον κανόνα τερματισμού ειδάλλως πηγαίνω στο βήμα 2.

Οι περισσότεροι από τους μη ιεραρχικούς αλγορίθμους διαφέρουν σε σχέση με:

1) Την μέθοδο που χρησιμοποιήθηκε για την απόκτηση των αρχικών κεντροειδών ή των κομβικών σημείων και

2) Τον κανόνα που χρησιμοποιήθηκε για την ανακατανομή των παρατηρήσεων.

Κάποιες από τις πιο βασικές μεθόδους για την απόκτηση των αρχικών κεντροειδών ή των κομβικών σημείων είναι:

- Επιλέγουμε τις k πρώτες παρατηρήσεις με μη ελλιπή δεδομένα σαν centroids ή seed points για αρχικά cluster.
- Επιλέγουμε τυχαία, k ,μη ελλιπής παρατηρήσεις σαν κέντρα clusters ή κομβικά σημεία.

Κάποιες από τις πιο βασικές μεθόδους για την ανακατανομή των παρατηρήσεων είναι:

- Υπολογίζουμε το κέντρο κάθε cluster(centroid) και αναθέτουμε εκ νέου τα αντικείμενα σ εκείνο το cluster με το πλησιέστερο centroid. Τα centroid δεν ενημερώνονται καθώς αναθέτουμε κάθε παρατήρηση στα k cluster. Αυτά υπολογίζονται εκ νέου αφού έχει γίνει η ανάθεση όλων των παρατηρήσεων. Αν η μεταβολή στα centroid των cluster είναι μεγαλύτερη από ένα κριτήριο σύγκλισης που ορίστηκε τα centroid επαναπροσδιορίζονται. Η διαδικασία ανακατανομής συνεχίζεται έως ότου η μεταβολή των centroid να είναι μικρότερη από το καθορισμένο κριτήριο σύγκλισης.
- Τοποθετούμε εκ νέου παρατηρήσεις έτσι ώστε να ελαχιστοποιείται κάποιο στατιστικό κριτήριο. Αυτές οι μέθοδοι είναι γνωστές και ως hill-climbing.

Από τις μη ιεραρχικές μεθόδους θα παρουσιασθεί η πιο ευρέως γνωστή η λεγόμενη K-means. Άλλη μέθοδος είναι η μέθοδος k-medoids η οποία κάνει χρήση κέντρου όχι ένα υπολογισμένο σημείο όπως η K-means αλλά ενός υπαρκτού σημείου δεδομένων.

3.1 Μέθοδος K-means

Η μέθοδος προτάθηκε από τον Macqeen (1967) και είναι η πιο διαδεδομένη μη ιεραρχική μέθοδος. Στόχος της είναι να κατανείμει ένα σύνολο αντικειμένων σε ένα προκαθορισμένο αριθμό συστάδων με τρόπο τέτοιο που να αυξάνει την ομοιότητα εντός των συστάδων. Ο αλγόριθμος περιλαμβάνει μια επαναληπτική διαδικασία όπου σε κάθε επανάληψη υπολογίζεται το κέντρο της συστάδας. Τα αντικείμενα εντάσσονται στη συστάδα με το πλησιέστερο κέντρο.
Αναλυτικά ο αλγόριθμος της K-means μεθόδου:
Βήμα 1: Διαμερίζουμε τα στοιχεία σε k αρχικά cluster και υπολογίζουμε το κέντρο του κάθε cluster(centroid).
Βήμα 2: Ανατρέχουμε εντός της λίστας των δεδομένων, αναθέτοντας κάθε δεδομένο στο cluster που έχει το κοντινότερο centroid. (Η απόσταση υπολογίζεται συνήθως με την ευκλείδεια απόσταση και ισούται με:

$$\text{mi}=\frac{1}{Mi}\sum_{j=1}^{Mi} xj$$

, όπου Mi το πλήθος των αντικειμένων της συστάδας i και mi το υπολογιζόμενο κέντρο).
Αφού αναθέσουμε όλα τα δεδομένα επαναυπολογίζουμε τις θέσεις των centroid για το cluster που λαμβάνει το νέο στοιχείο και το cluster που χάνει το στοιχείο.

Βήμα 3: Επαναλαμβάνουμε το βήμα 2 έως ότου δεν πραγματοποιηθούν όλες οι ανακατατάξεις και τα centroid δεν κινούνται.

Παρά την έναρξη με μια διαμέριση όλων των στοιχείων σε k αρχικές ομάδες στο βήμα 1, θα μπορούσε να διευκρινιστούν τα αρχικά centroids και μετά να συνεχίσουμε στο βήμα 2, δηλαδή , Βήμα 1: Τοποθετούμε k σημεία στο χώρο των δεδομένων που θα ομαδοποιηθούν τα οποία αντιπροσωπεύουν τα αρχικά centroids.

Ένας από τους βασικούς στόχους του αλγορίθμου είναι να καταφέρει να ελαχιστοποιήσει τη συνάρτηση τετραγωνικού λάθους που ισούται :

$$V=\sum_{i=1}^{k}\sum_{xj\in Ci}(xj-mi)^2,$$

Όπου ci οι συστάδες, x τα αντικείμενα και mi τα κέντρα της συστάδας i.

3.1.1 Ομαδοποίηση με τη Μέθοδο K-means

Υποθέτουμε ότι μετράμε 2 μεταβλητές x1,x2 για κάθε ένα από τα 4 στοιχεία A,B,C,D. Τα δεδομένα δίνονται στον ακόλουθο πίνακα:

Στοιχεία	**Παρατήρηση x1**	**Παρατήρηση x2**
A	5	3
B	-1	1
C	1	-2
D	-3	-2

Επιλέγω να διαιρέσω αυτά τα στοιχεία σε έστω k=2 cluster έτσι ώστε τα στοιχεία ενός cluster να είναι πιο όμοια μεταξύ τους από ότι σε διαφορετικό cluster.

Για να εφαρμόσουμε την μέθοδο k=2-means διαιρούμε αυθαίρετα τα στοιχεία σε 2 cluster, τα (AB), (CD) και υπολογίζουμε τις συντεταγμένες (x1,x2) του centroid του κάθε cluster.

Άρα στο βήμα 1 έχουμε,

cluster	**x1**	**x2**
(AB)	$\frac{5+(-1)}{2}=2$	$\frac{3+(1)}{2}=2$
(CD)	$\frac{1+(-3)}{2}=-1$	$\frac{-2+(-2)}{2}=-2$

Στο βήμα 2 υπολογίζουμε την ευκλείδεια απόσταση κάθε στοιχείου από την ομάδα των centroid και αναθέτουμε ξανά κάθε στοιχείο στην πλησιέστερη σε αυτό ομάδα. Αν ένα στοιχείο μετακινείται από την αρχική κατάσταση, τα κέντρα των cluster πρέπει να ενημερωθούν πριν προχωρήσουμε.

Η i- συντεταγμένη του centroid, i=1,2,…p ενημερώνεται από τους τύπους:

$$xi(new)=\frac{nxi+xji}{n+1}$$, αν το j-στοιχείο προστίθεται σε μια ομάδα

$$xi(new)=\frac{nxi-xji}{n-1}$$, αν το j-στοιχείο αφαιρείται από μια ομάδα

Εδώ n ο αριθμός στοιχείων στην παλιά ομάδα με centroid τα x1,x2. Θεωρούμε τα αρχικά cluster (AB), (CD). Οι συντεταγμένες των centroid είναι (2,2) και (-1,-2) αντίστοιχα.

Έστω το στοιχείο A(5,3) μετακινείται στην ομάδα (CD). Οι νέες ομάδες είναι (B) και (ACD) και με τα ενημερωμένα centroid:

Ομάδα (B): $x1(new)=\frac{2(2)-5}{2-1}=-1$ και $x2(new)=\frac{2(2)-3}{2-1}=1$

Ομάδα (ACD): $x1(new)=\frac{2(-1)+5}{2+1}=1$ και $x2(new)=\frac{2(-2)+3}{2+1}=0.33$

Επιστρέφουμε στις αρχικές ομάδες του βήματος 1 και υπολογίζουμε τις τετραγωνικές αποστάσεις:

d^2(A, (AB))=$(5-2)^2+(3-2)^2=10$

d^2(A, (CD))=$(5+1)^2+(3+2)^2=61$, αν το A δεν μετακινείται

d^2(A, (B)=$(5+1)^2+(3-1)^2=40$

d^2(A, (ACD))=$(5-1)^2+(3-0.33)^2=27.09$, αν το A μετακινείται στην (CD)

Αφού το A βρίσκεται πλησιέστερα στο κέντρο του (AB) απ’ ότι στο κέντρο του (ACD) δεν τοποθετείται εκ νέου.

Συνεχίζουμε με το B:

d^2(B, (AB))=$(-1-2)^2+(1-2)^2=10$

d^2(B, (CD))=$(-1+1)^2+(1+2)^2=9$, αν το B δεν μετακινείται

d^2(B, (A))=$(-1-5)^2+(1-3)^2=40$

d^2(B, (BCD))=$(-1+1)^2+(1+1)^2=4$, αν το B μετακινείται στην ομάδα (CD)

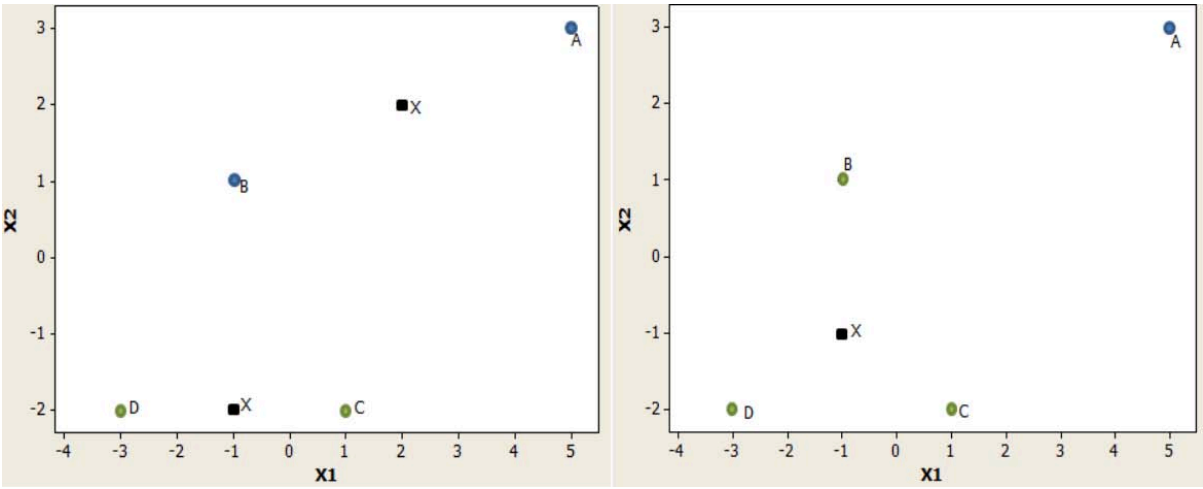
Αφού το B βρίσκεται πλησιέστερα στο κέντρο του (BCD) απ’ ότι στο κέντρο του (AB), το B μεταφέρεται στην ομάδα (CD).

Τώρα έχουμε (A) με (5,3) και (BCD) με (-1,-1).

Συνεχίζοντας όμοια με το C βλέπουμε πώς το C δεν μετακινείται και τελικά βρίσκουμε πως δεν μπορούν να υπάρξουν περισσότερες ανακατατάξεις και τα τελικά cluster k=2 είναι (A) και (BCD) με το άθροισμα των τετραγώνων στο εσωτερικό του κάθε cluster να είναι:

(A): 0 και (BCD): 4+5+5=14.

Παρακάτω φαίνεται σχηματική αναπαράσταση του k-means όπου αριστερά φαίνεται η αρχική τυχαία διαμέριση που επιλέξαμε εμείς και δεξιά το τελικό αποτέλεσμα.



3.2 Πλεονεκτήματα και Μειονεκτήματα Μεθόδου K-means

Ο αλγόριθμος K-means διαθέτει τα παρακάτω πλεονεκτήματα:

- Είναι απλός και κατανοητός.
- Τα αντικείμενα μοιράζονται σε συστάδες με αυτόματο τρόπο.
- Είναι αρκετά γρήγορος. Για τον λόγο αυτό είναι περισσότερο κατάλληλος για ομαδοποίηση μεγάλων συνόλων δεδομένων σε σχέση με άλλες μεθόδους.
-

Ωστόσο παρουσιάζονται και κάποια μειονεκτήματα:

- Ο αριθμός των συστάδων πρέπει να καθοριστεί από τον χρήστη, έτσι ίσως απαιτείται σε κάποιες περιπτώσεις να την εφαρμόσει ένας έμπειρος χρήστης
- Το τελικό αποτέλεσμα εξαρτάται σε μεγάλο βαθμό από την αρχική επιλογή των κέντρων. Σε αντίθετη περίπτωση μπορεί να οδηγηθούμε σε σημαντικά διαφορετικές συστάδες.
- Είναι πολύ ευαίσθητος στη ύπαρξη ακραίων τιμών(outliers). Λίγα αντικείμενα με πολύ μεγάλες τιμές μπορεί να επηρεάσουν των υπολογισμό των νέων κέντρων και συνεπώς την διαμόρφωση των τελικών συστάδων.
- Έχει την τάση να δημιουργεί σφαιρικές ή ίσου μεγέθους συστάδες, οπότε δεν είναι κατάλληλος για συστάδες με περίπλοκα σχήματα ή διαφορετικού μεγέθους.

ΚΕΦΑΛΑΙΟ 4

ΣΥΜΠΕΡΑΣΜΑΤΑ

Όπως περιγράψαμε προηγουμένως οι μη ιεραρχικές τεχνικές ομαδοποίησης απαιτούν γνώση για τον αριθμό των cluster σε αντίθεση με τις ιεραρχικές όπου δεν απαιτείται εκ των προτέρων γνώση του αριθμού των cluster ή της αρχικής διαμέρισης. Αυτό είναι το πλεονέκτημα των ιεραρχικών μεθόδων έναντι των μη ιεραρχικών καθώς επίσης πως είναι υπολογιστικά γρηγορότερες. Όμως οι ιεραρχικές έχουν το μειονέκτημα ότι άπαξ και μια παρατήρηση ανατεθεί σε ένα cluster δεν μπορεί να ανατεθεί σε κάποιο άλλο στην συνέχεια. Οι αλγόριθμοι μη ιεραρχικής ομαδοποίησης είναι πολύ ευαίσθητοι στην αρχική διαμέριση. Σε περιπτώσεις που ο χρήστης επιλέγει το k καλή ιδέα είναι να ξανατρέξουμε τον αλγόριθμο για διάφορες επιλογές του k για να πάρουμε ένα ασφαλή αποτέλεσμα όσο αφορά την τελική σύσταση των συστάδων. Αποτέλεσμα μελετών έχουν δείξει πως η μέθοδος k-means αλλά και άλλοι μη ιεραρχικές μέθοδοι έχουν χαμηλή απόδοση όταν χρησιμοποιούν τυχαίες διαμερίσεις. Όμως η απόδοσή τους είναι ανώτερη όταν χρησιμοποιούνται αποτελέσματα από ιεραρχικές μεθόδους για τον σχηματισμό της αρχικής διαμέρισης. Επομένως συνίσταται ότι για μη ιεραρχικές μεθόδους θα χρησιμοποιούσε κανείς μια εκ των προτέρων αρχική διαμέριση ή λύση των cluster. Με άλλα λόγια οι ιεραρχικές και μη ιεραρχικές θα μπορούσαν να θεωρηθούν ως συμπληρωματικές και όχι ως ανταγωνιστικές. Κατά συνέπεια οι ιεραρχικές μέθοδοι χρησιμοποιούνται κάποιες φορές με μια διερευνητική έννοια και η λύση τους υποβάλλεται σε μια μη ιεραρχική μέθοδο για την περαιτέρω βελτίωση.
Όμως ποια από τα 2 είδη είναι καλύτερο?? Έπειτα αφού ο ερευνητής επιλέξει μια από τις 2 τεχνικές(ιεραρχική και μη ιεραρχική) ποια μέθοδο(π.χ. single linkage ή average linkage) πρέπει να επιλέξει?? Προφανώς η συζήτηση εξαρτάται από το αντικείμενο της μελέτης και τις ιδιότητες των διαφόρων αλγορίθμων.

ΑΝΑΦΟΡΕΣ

[1] Richard A. Johnson and Wichern (1992), "Applied Multivariate Statistical Analysis", 3rd edition, New Jersey: Prentice- Hall International, Inc.

[2] D. J. Hand, "Discrimination and Classification", John Wiley & Sons.

[3] Osama Abu Abbas, comparison between data clustering algorithms, the international Arab journal, july 2008

[4] Καράγεωργα Ισμήνη, Ανάλυση Συστάδων,2012

[5] Ζιντζαράς Ηλίας, 2016-2017, course advance statistics.